An Empirical Study on the Use of Visual Explanation in Kidney Cancer Detection

Masaya Takahashi[a], Yoshitaka Kameya[*a], Keiichi Yamada[a], Kazuhiro Hotta[a], Tomoichi Takahashi[a], Naoto Sassa[b,c], Shingo Iwano[b] and Tokunori Yamamoto[b]

[a]Grad. Sch. of Science and Technology, Meijo University, 1-501 Shiogama-guchi, Tenpaku-ku, Nagoya, Japan, 468-8502; [b]Grad. Sch. of Medicine, Nagoya University, 65 Tsurumai-cho, Showa-ku, Nagoya, Japan, 466-8560; [c]School of Medicine, Aichi Medical University, 1-1 Yazakokarimata, Nagakute, Japan, 480-1195

ABSTRACT

In order to detect kidney cancer automatically from abdominal UCT (unenhanced CT) or CECT (contrast-enhanced CT) images at an early stage, a promising approach is to use deep learning techniques with convolutional neural networks (CNNs). However, there still seem to be several challenges in detection of kidney cancer. For example, it is necessary to cope with the wide variety of abdominal CT images. In this paper, as an empirical study, we attempt to construct a CNN that detects kidney cancer well from abdominal CT images, with a special focus on how visual explanations produced by Gradient-weighted Class Activation Mapping (Grad-CAM) help us to construct such a CNN.

Keywords: Kidney Cancer, Computer Tomography (CT), Convolutional Neural Network (CNN), Visual Ex- planation, Gradient-weighted Class Activation Mapping (Grad-CAM)

1. INTRODUCTION

One major way for detecting kidney cancer from abdominal computer tomography (CT) images is to use contrast agents, which enhance the contrast among tissues in a kidney. CT images obtained under an administration of contrast agents are called contrast-enhanced CT (CECT) images, and medical experts often diagnose kidneys using a couple of CECT images taken at predetermined time intervals.[1, 2] For some patients, on the other hand, unenhanced CT (UCT) is highly preferred if they have allergy for contrast agents or injecting contrast agents worsens their renal function. In this study, aiming to detect kidney cancer automatically from UCT or CECT images at an early stage, we take a deep learning approach using convolutional neural networks (CNNs).

Although CNNs have shown remarkable predictive performance for image-related tasks, there still seem to be several challenges in detection of kidney cancer from CT images. First, unlike well-known benchmark datasets for image classification such as ImageNet, abdominal CT images of an individual patient are basically very similar, and a CNN should capture the patient's kidney and a cancer in it in a precise way. Especially, it is known to be difficult to detect a kidney cancer in a kidney from UCT images, since the cancer often has a texture quite similar to that of the other tissues in the kidney.[3] Second, since abdominal organs locate at various positions depending on individual patients, it is also necessary to tackle with the wide variety of abdominal CT images. Lastly, CNNs are said to be black-box models, and from both clinical and engineering points of view, we need some explanation method for understanding the reasons why their decisions have been made.[4] One well-known visual explanation method is Gradient-weighted Class Activation Mapping (Grad-CAM),[5] which can highlight some regions in the input image relevant to the CNN's decision. In this paper, as an empirical study, we attempt to construct a CNN that detects kidney cancer well from abdominal CT images, with a special focus on how visual explanations produced by Grad-CAM help us to construct such a CNN.

*ykameya@meijo-u.ac.jp; phone: +81 52 838 2567

The rest of this paper is organized as follows. First, we introduce some background notions in Section 2. Then, the methods and the experimental settings adopted in this study are described in Section 3. Section 4 discusses the relations between the detection accuracy of CNNs and visual explanations produced by Grad-CAM, and Section 5 concludes the paper. This study was approved by the Ethics Committees at Nagoya University and Meijo University.

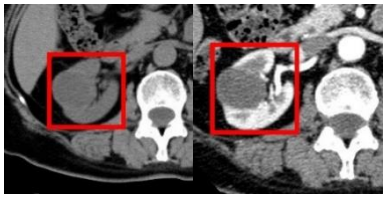


Figure 1. An UCT image (left) and the corresponding CECT image (right).

2. BACKGROUND

2.1 Contrast-enhanced CT and endophytic kidney cancer

As mentioned above, the texture of kidney cancer and that of normal tissues are often similar in UCT images, and thus it is often difficult to detect kidney cancer from UCT images. What is worse is that endophytic kidney cancer grows inside a kidney and does not deform the contour of the kidney.[3] CECT images mitigate this problem by presenting different textures between cancer and normal tissues after the injection of a contrast agent. A typical example is shown in Fig. 1, where the left image is an UCT image, and the right one is the corresponding CECT image. In these images, kidneys are surrounded by red rectangles, and one would see that kidney cancer can be found more clearly in the CECT image. In this study, we observe the detection accuracy of CNNs trained with UCT images and those trained with CECT images, and pay an additional attention to their detection accuracy for endophytic kidney cancer.

2.2 Convolutional neural networks

In this study, we attempt to detect kidney cancer by classifying abdominal CT images into those containing the cancer and those not containing the cancer, and use a CNN for such classification. CNNs are typically comprised of convolutional layers, pooling layers, and fully-connected layers, and known for its high predictive performance. In medical image processing, CNNs have been used for detecting kidney cancer,6 and for classifying kidney cancer into three subtypes.[1] In the former study, to achieve patient-wise detection of cancer, the authors proposed a CNN which takes as input a collage of multiple abdominal CT images of a patient, and predicts the presence/absence of the cancer for the patient. In this study, as described later, we take a simpler approach to patient-wise detection, exploiting the class labels given to individual CT images.

2.3 Visual explanation by Grad-CAM

One practical problem in using a deep neural network is the lack of explainability for its prediction. Recently, in image classification, dozens of visual explanation methods that indicate the crucial regions in the input image for a CNN's prediction have been proposed.[4] Grad-CAM is a well-known visual explanation method, which is said to be class-discriminative in the sense that a visual explanation for a particular target class (the presence/absence of cancer, in this study) *exclusively* highlights relevant regions in the input image.[5] In medical image processing, Philbrick et al. used a couple of visual explanation methods including Grad-CAM for CNNs that identify the contrast enhancement phase of an input CT image.[2] Seto et al. applied SmoothGrad, another well-known visual explanation method, to their CNN that detects esophageal cancer from chest CT images.[7] In this study, we will discuss more extensively how visual explanations change under various experimental settings, and help us to improve the detection accuracy of CNNs.

3. METHODS AND SETTINGS

3.1 Datasets

In this study, we worked with a dataset comprised of UCT images and that of CECT images, both of which have been collected at Nagoya University Hospital. The CECT images we used were the ones taken 30 seconds after the injection of

a contrast agent. In these datasets, each CT image was associated with a class label indicating whether a kidney cancer is present or absent, based on the axial positions of the top-end and the bottom-end of a kidney and a kidney cancer (if it exists) suggested by a medical expert. Then, we split each original dataset randomly into the training dataset, the validation dataset, and the evaluation dataset, with keeping the class distribution. For simplicity, we just focused on the right kidney. Like the CT images shown in Fig. 1, from each original CT image of size (512, 512), we cropped out the area of size (256, 256) that ranges [40, 296) horizontally and [186, 442) vertically, in order to reduce unwanted influence from the other organs around a kidney. The numbers of patients in split datasets are shown in Table 1. Note here that the 'Both' column indicates the number of both patients having endophytic cancer and those having exophytic cancer, while the 'Endophytic' column indicates the number of patients having endophytic cancer only. Also note that the figures in parentheses in Table 1 are the numbers of available UCT and CECT images, where the numbers of CT images not containing cancer of the patients having cancer are included in the 'Cancer: Absent' columns. In the rest of this paper, we think of a patient or a CT image having cancer as positive, and as negative otherwise.

Table 1. The number of patients in each dataset split from the original dataset.

Dataset	UCT				CECT			
	Cancer: Present		Cancer: Absent		Cancer: Present		Cancer: Absent	
	Both	Endophytic			Both	Endophytic		
Training	45 (228)	15 (71)	29	(1,269)	43 (188)	13 (52)	29	(1,127)
Validation	9 (33)	3 (10)	6	(274)	9 (28)	3 (9)	5	(236)
Evaluation	10 (47)	4 (17)	5	(253)	10 (36)	4 (12)	5	(221)

3.2 Finding the correspondence between UCT images and CECT images

Actually the positions of the top-end and the bottom-end of kidneys and kidney cancer were only suggested for CECT images, and we needed to find the correspondence between UCT images and CECT images, like the images shown in Fig. 1. An intricate problem here is that the absolute/relative positions of individual abdominal organs constantly change even in the same patient, and UCT images and CECT images have been taken at different intervals (5 mm and 1 mm, respectively) on the axial axis under different conditions of a patient. The class labels of UCT images can therefore be incorrect. To avoid this, we relabeled each UCT image with the class of the most similar CECT image under the cosine similarity.

3.3 Data augmentation

As stated in the introduction, abdominal organs locate at various positions depending on individual patients, and thus we have a wide variety of abdominal CT images. This urges us to construct a CNN with higher generalization performance. One general technique for acquiring generalization performance is data augmentation, and in this study, we augmented the training dataset with additional CT images created by the transformations listed in Table 2. The degree of each transformation was configured so that kidneys and kidney cancer do not disappear from the transformed image. The training dataset then turned to be 25 times larger than the original. Both original and transformed images were finally cropped with size (224, 224) at the center, and given to the CNN.

Table 2. Details of data augmentation conducted in this study.

Transformation	Details	#augmented
Shift	$\{0 \sim 8, 8 \sim 16\}$ pixels upward, downward, to the left and to the right	8
Rotation	$\{0 \sim 8, 8 \sim 16, 16 \sim 24, 24 \sim 32\}$ degree clockwise and anti-clockwise	8
Shear transformation	$\{-15 \sim 0, 0 \sim 15\}$ degree horizontally and vertically	4
Zooming-in	By a factor of $\{1 \sim 272/256, 272/256 \sim 288/256\}$	2
Zooming-out	By a factor of $\{224/256 \sim 240/256, 240/256 \sim 1\}$	2

3.4 Masking the other organs

During the experiments, we considered that CNNs suffer from unwanted influence from the other organs around a kidney. So, as a trial, a couple of non-experts masked such organs manually as shown in Fig. 2, where a kidney is surrounded by a red square in the left image. The effect of masking will be discussed in Section 4 based on detection accuracy and visual explanations. Table 3 shows the numbers of patients (and the number of CT images) in split datasets, where we just deleted masked images of apparently low quality.

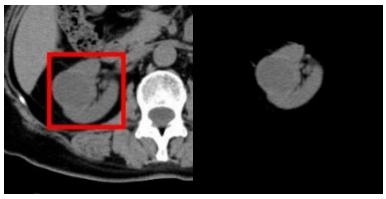


Figure 2. An original UCT image (left) and the masked one (right).

Table 3. The number of patients in each dataset comprised of masked images.

Dataset	UCT						CECT					
	Cancer: Present				Cancer: Absent		Cancer: Present				Cancer: Absent	
	Both		Endophytic				Both		Endophytic			
Training	45	(223)	15	(71)	29	(1,238)	43	(187)	13	(52)	29	(1,113)
Validation	9	(33)	3	(10)	5	(265)	9	(27)	3	(9)	5	(234)
Evaluation	10	(43)	4	(17)	5	(240)	10	(36)	4	(12)	5	(221)

3.5 Coping with class imbalance

Also during the experiments, we encountered a kind of class imbalance problem.[8] To be more specific, in Table 1, the training dataset comprised of UCT images contains 223 positive images and 1,238 negative images, and by CNNs trained on this training dataset, an excessive number of images in the evaluation dataset were classified as negative. We tackled with this problem in two ways. The former is to give a weight N^{-}_{train} to positive patients and N^{+}_{train} to negative patients in the cross-entropy loss function, where N^{+}_{train} (resp. N^{-}_{train}) is the total number of positive (resp. negative) CT images in the training dataset. The latter is to construct a balanced training dataset by selecting the most similar negative CT image for each positive CT image.[7] Hereafter, the former method is referred to as instance weighting (IW), and the latter as negative example selection (NES).

3.6 Training CNNs

The CNN we used in this study is a variant of VGG-16,[9] which has 13 convolutional layers and three fully-connected layers together with batch normalization. We performed transfer learning where the weights in convolutional layers were pre-trained with ImageNet and those in fully-connected layers were trained by ourselves with the training dataset. In a preliminary experiment, fully-trained models did not perform as well as the pre-trained model. The training was conducted by AdaGrad with mini-batch size 32, initial learning rate 10^{-5}, and dropout rate 0.5. We finally chose the CNN having achieved the highest accuracy over the validation dataset after 50 epochs.

3.7 Image- and patient-wise detection

In this study, we detect kidney cancer of a patient in two steps. In the first step, called image-wise detection, we classify each abdominal CT image of the patient into the one containing cancer and the one not containing cancer. Then, in the second step, called patient-wise detection, we combine the results of image-wise detection for the patient's CT images. Of course, only patient-wise detection is meaningful in a clinical sense, but it is also useful to inspect the results of image-wise detection to see how well the CNN has been trained.

More specifically, let us consider a set $X(p)$ of abdominal CT images of a patient p. Then, in image-wise detection for each abdominal CT image $x \in X(p)$, the softmax layer of the trained CNN outputs the confidence $c = P(\textit{positive} \mid x)$ that the image x contains cancer. Now we predict that the image x contains cancer (x is

positive) if $c \geq \theta_{\text{image}}$, where θ_{image} is the decision threshold for image-wise detection, and that x does not contain cancer (x is negative) otherwise. In patient-wise detection for a patient p, on the other hand, we just compute the maximum confidence $c^* = \max_{x \in X(p)} P(\textit{positive} \mid x)$ over all abdominal CT images of the patient p, and then predict that the patient p has cancer (p is positive) if $c^* \geq \theta_{\text{patient}}$, where θ_{patient} is the decision threshold for patient-wise detection, and that p does not have cancer (p is negative) otherwise. In this study, for simplicity, the decision thresholds θ_{image} and θ_{patient} were fixed at 0.5 and 0.95, respectively.

3.8 Running Grad-CAM

After image-wise detection, we run Grad-CAM to obtain its visual explanation. Our PyTorch implementation for Grad-CAM is based on the one described in https://www.noconote.work/entry/2019/01/12/231723, a technical blog post written in Japanese. To create such an explanation, Grad-CAM focuses on the feature maps in the final convolutional layer of the trained CNN, which is supposed to hold meaningful high-level features, and identifies relevant regions based on the weighted average of these feature maps. The class-discriminativity of Grad-CAM comes from a design where each weight is computed as the average gradient of activation for a target class. In this study, we considered that the target class is the class predicted by the trained CNN. The relevant regions are finally highlighted in the form of a heatmap which is superimposed on the target abdominal CT image.

Table 4. The results of image-wise detection.

Setting	TP	FN	TN	FP	Precision	Recall	F-score
UCT+Unmasked+IW	13 (3)	34 (14)	230	23	0.361 (0.115)	0.277 (0.177)	0.313 (0.140)
UCT+Unmasked+NES	32 (7)	15 (10)	137	116	0.216 (0.057)	0.681 (0.412)	0.328 (0.100)
UCT+Masked+NES	29 (9)	14 (5)	202	38	0.433 (0.192)	0.674 (0.643)	0.527 (0.295)
CECT+Unmasked+NES	6 (4)	30 (8)	205	16	0.273 (0.200)	0.167 (0.333)	0.207 (0.250)
CECT+Masked+NES	28 (10)	8 (2)	162	59	0.322 (0.145)	0.778 (0.833)	0.455 (0.247)

4. RESULTS AND DISCUSSION

4.1 Detection accuracy

We first show the detection accuracy of the trained CNNs in image-wise and patient-wise detection under various experimental settings. An experimental setting is a combination of three binary choices, i.e. {UCT, CECT} (whether we use UCT images or CECT images), {Unmasked, Masked} (whether we use unmasked images or masked images), and {IW, NES} (whether we conduct instance weighting or negative example selection to cope with class imbalance). In what follows, each setting is referred to like UCT+Unmasked+IW, and we focus on five settings UCT+Unmasked+IW, UCT+Unmasked+NES, UCT+Masked+NES, CECT+Unmasked+NES, and CECT+Masked+NES which produced relatively high detection accuracy. One may find that, among these five settings, a setting with lower (clinical, annotation, or computational) cost comes earlier.

4.1.1 Image-wise detection

The results of image-wise detection under the five settings above are shown in Table 4, where we abbreviate the number of true positives, false positives, true negatives, and false negatives as TP, FP, TN, and FN, respectively. Three evaluation metrics, i.e. precision, recall, and F-score, are computed w.r.t. the positive class, i.e. the presence of cancer. The figures in parentheses in Table 4 are the ones related to endophytic cancer. For example, in Table 4, precision w.r.t. the presence of endophytic cancer under the UCT+Unmasked+IW setting is computed as $3 / (3 + 23) = 0.115$.

We can say from Table 4 as follows. First, recall with negative example selection (NES) was higher than that with instance weighting (IW), at the cost of the increase in the number of false positives. Second, with masked images, all evaluation metrics were improved. This would imply that the influence from the other organs around the kidney is not ignorable. Third, while CECT images did not improve detection accuracy as expected, the decrease of evaluation metrics

for endophytic cancer was not observed with CECT images. Lastly, as might be expected, we achieved the highest recall under the most costly setting, i.e. CECT+Masked+NES.

4.1.2 Patient-wise detection

The results of patient-wise detection under the selected five settings are shown in Table 5, where all evaluation metrics were higher than those in image-wise detection, even for endophytic cancer. In particular, we achieved quite high recall due to our way for patient-wise detection that exploits the maximum confidence over all CT images of a patient. It should also be noted that the evaluation metrics were sensitive to the threshold $\theta_{patient}$ in patient-wise detection.

Table 5. The results of patient-wise detection.

Setting	TP	FN	TN	FP	Precision	Recall	F-score
UCT+Unmasked+IW	3 (1)	7 (3)	4	1	0.750 (0.500)	0.300 (0.250)	0.429 (0.333)
UCT+Unmasked+NES	10 (4)	0 (0)	1	4	0.714 (0.500)	1.000 (1.000)	0.833 (0.667)
UCT+Masked+NES	10 (4)	0 (0)	2	3	0.769 (0.571)	1.000 (1.000)	0.870 (0.727)
CECT+Unmasked+NES	3 (2)	7 (2)	4	1	0.750 (0.667)	0.300 (0.500)	0.429 (0.571)
CECT+Masked+NES	10 (4)	0 (0)	2	3	0.769 (0.571)	1.000 (1.000)	0.870 (0.727)

4.2 Visual Explanation

From now on, we show several examples of visual explanations produced by Grad-CAM under each of the selected five settings, and discuss the relations between the accuracy of image-wise detection and visual explanations.

4.2.1 UCT+Unmasked+IW

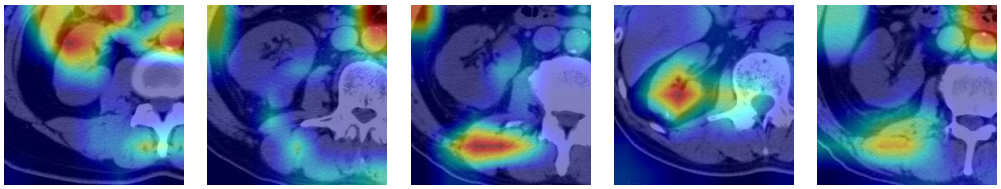
Fig. 3 shows several visual explanations in the form of heatmaps under the UCT+Unmasked+IW setting. One may find that, for the CT images in Figs. 3 (a) and 3 (d), the trained CNN surely looked at the kidney, but it often performed classification based on the other organs as exhibited by Figs. 3 (b), 3 (c), and 3 (e). Overall, the kidney did not frequently overlap with the regions relevant to the CNN's decision.

4.2.2 UCT+Unmasked+NES

Fig. 4 shows visual explanations under the UCT+Unmasked+NES setting. For the images in Figs. 4 (a), 4 (b), and 4 (d), the trained CNN decided the presence/absence of cancer by looking at the kidney, but for the images in Figs. 4 (c) and 4 (e), it relied on the regions around the kidney. Overall, thanks to negative example selection, the trained CNN turned to be sensitive to the differences inside the kidney, not in the surrounding organs, and it achieved a better detection accuracy than that under the UCT+Unmasked+IW setting.

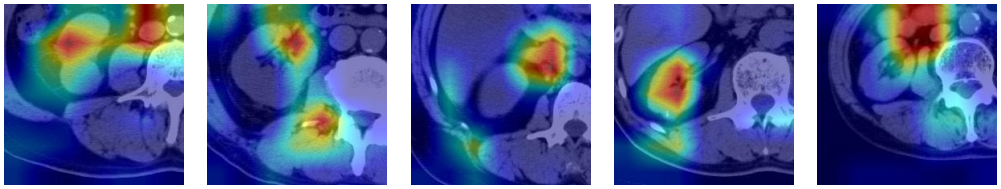
4.2.3 UCT+Masked+NES

Fig. 5 shows visual explanations under the UCT+Masked+NES setting. For the images in Figs. 5 (a), 5 (b), 5 (c), and 5 (e), which were classified as positive, it is obvious that the trained CNN looked only at the kidney. As Table 4 says, the accuracy of image-wise detection was much improved under this setting, and thus reducing irrelevant information in the input image is particularly important. It should be noted however that the trained CNN's decision was sensitive to the contour of the kidney, and eventually, to the quality of masking.



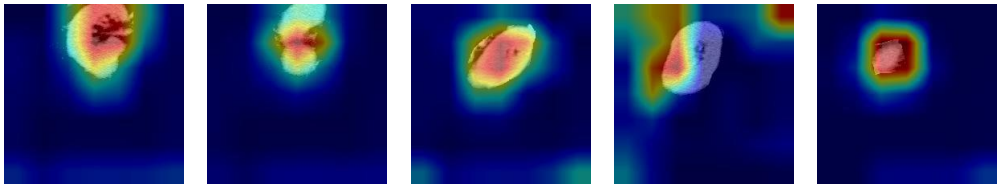
(a) (b) (c) (d) (e)

Figure 3. (a)–(c) Heatmaps in true positive cases under the UCT+Unmasked+IW setting, (d) a heatmap in false negative cases, and (e) a heatmap in false positive cases.



(a) (b) (c) (d) (e)

Figure 4. (a)–(c) Heatmaps in true positive cases under the UCT+Unmasked+NES setting, (d) a heatmap in false negative cases, and (e) a heatmap in false positive cases.

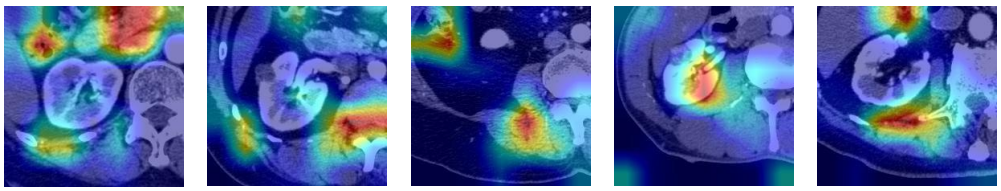


(a) (b) (c) (d) (e)

Figure 5. (a)–(c) Heatmaps in true positive cases under the UCT+Masked+NES setting, (d) a heatmap in false negative cases, and (e) a heatmap in false positive cases.

4.2.4 CECT+Unmasked+NES

Fig. 6 shows visual explanations under the CECT+Unmasked+NES setting. For the images in Figs. 6 (a), 6 (b), 6 (c), and 6 (e), which were classified as positive, the trained CNN looked at the surrounding organs. On the other hand, for the image in Fig. 6 (d), which was classified as negative, the trained CNN looked at normal tissues in the kidney, which were mostly whitened in CECT images. One hypothesis for the latter is that the decision by the trained CNN may have been made based on the texture of a kidney rather than its contour, and hence there would be a room for improvement of the CNN's structure or its training process.

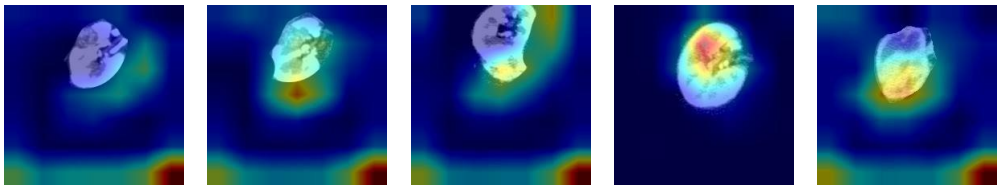


(a) (b) (c) (d) (e)

Figure 6. (a)–(c) Heatmaps in true positive cases under the CECT+Unmasked+NES setting, (d) a heatmap in false negative cases, and (e) a heatmap in false positive cases.

4.2.5 CECT+Masked+NES

Lastly, Fig. 7 shows visual explanations under the CECT+Masked+NES setting. For the images in Figs. 7 (a), 7 (b), 7 (c), and 7 (e), which were classified as positive, the focus of the trained CNN is primarily on the kidney. Similarly to the case under the UCT+Masked+NES setting, we can say that masking is effective in reducing the unwanted influence, while its quality is crucial. In addition, contrastingly with the case under the CECT+Unmasked+NES setting, the kidney was surely taken into account.



(a) (b) (c) (d) (e)

Figure 7. (a)–(c) Heatmaps in true positive cases under the CECT+Masked+NES setting, (d) a heatmap in false negative cases, and (e) a heatmap in false positive cases.

5. CONCLUSION

In this paper, as an empirical study, we constructed a CNN that detects kidney cancer well from abdominal CT images, and discussed the relations between the CNN's detection accuracy and visual explanations produced by Grad-CAM. With a help from visual explanations that highlight the regions relevant to the CNN's decision, we improved the detection accuracy of the CNN by elaborating the dataset. For example, reducing the unwanted influence from the organs around a kidney is found to be the most effective, and selection of negative examples is a promising vehicle that only requires a computational cost. As is often said,10 for the decision made by a deep neural network, visual explanation is inherently just an approximation of the true explanation. However, at least in this study, it often gave us clues for improving predictive models and planning experiments. We would also like to add that visual explanation played a key role in discussion with medical experts.

In future, we will work for further improvement of the CNN's detection accuracy. We are planning to extend the datasets of UCT and CECT images, and to refine the method for patient-wise detection. In patient-wise detection, decision thresholds can be opt based on the validation dataset, or more sophisticatedly, following Seto et al., we can combine high-level features extracted from multiple CT images of a patient as an input of a recurrent neural network,7 provided more CT images are available. It is also interesting to introduce a recent visual explanation method equipped with class-discriminativity and high resolution at the same time.[11]

REFERENCES

[1] S. Han, S. I. Hwang, and H. J. Lee, "The classification of renal cancer in 3-phase CT images using a deep learning method," *J. of Digital Imaging* **32**(4), pp. 638–643, 2019.

[2] K. A. Philbrick, K. Yoshida, D. Inoue, Z. Akkus, T. L. Kline, A. D. Weston, P. Korfiatis, N. Takahashi, and B. J. Erickson, "What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images," *American J. of Roentgenology* **221**(6), pp. 1184–1193, 2018.

[3] S. D. O'Connor, S. G. Silverman, L. R. Cochon, and R. K. Khorasani, "Renal cancer at unenhanced CT: Imaging features, detection rates, and outcomes," *Abdominal Radiology* **43**(7), pp. 1756–1763, 2018.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Predreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys* **51**(5), pp. 93:1–93:42, 2018.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV-17)*, 2017.

[6] M. A. Hussain, A. Amir-Khalili, G. Hamarneh, and R. Abugharbieh, "Collage CNN for renal cell carcinoma detection from CT," in *Proc. of the 8th Int'l Workshop on Machine Learning in Medical Imaging (MLMI-17)*, pp. 229–237, 2017.

[7] T. Seto, M. Takeuchi, M. Hashimoto, Y. Ito, N. Ichihara, H. Kawakubo, Y. Kitagawa, H. Miyata, M. Jinzaki, and Y. Sakakibara, "Classification of esophageal cancer CT images using deep learning," in *Proc. of The 33rd Annual Conf. of the Japanese Society for Artificial Intelligence (JSAI-19)*, 2019. In Japanese.

[8] G. M. Weiss, "Foundations of imbalanced learning," in *Imbalanced Learning*, H. He and Y. Ma, eds., pp. 13– 42, Wiley, 2013.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of the 3rd Int'l Conf. on Learning Representation (ICLR-15)*, 2015.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence* **1**, pp. 206–215, 2019.

[11] H. Tsunakawa, Y. Kameya, H. Lee, Y. Shinya, and N. Mitsumoto, "Contrastive relevance propagation for interpreting predictions by a single-shot object detector," in *Proc. of the 2019 Int'l Joint Conf. on Neural Networks (IJCNN-2019)*, 2019.